D548–D551 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh093

EICO (Expression-based Imprint Candidate Organizer): finding disease-related imprinted genes

Itoshi Nikaido[1,2,3], Chika Saito[3], Akiko Wakamoto[3], Yasuhiro Tomaru[3], Takahiro Arakawa[3], Yoshihide Hayashizaki[1,3] and Yasushi Okazaki[2,3,*]

[1]Division of Genomic Information Resource Exploration, Science of Biological Supramolecular Systems, Yokohama City University, Graduate School of Integrated Science, Yokohama, Kanagawa 230-0045, Japan, [2]Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical School, Saitama 350-1241, Japan and [3]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan

Received August 13, 2003; Revised and Accepted October 7, 2003

ABSTRACT

We have developed an integrated database that is specialized for the study of imprinted disease genes. The database contains novel candidate imprinted genes identified by the RIKEN full-length mouse cDNA microarray study, information on validated single nucleotide polymorphisms (SNPs) to confirm imprinting using reciprocal mouse crosses and the predicted physical position of imprinting-related disease loci in the mouse and human genomes. It has two user-friendly search interfaces: the SNP-central view (MuSCAT: MoUse SNP CATalog) and the candidate gene-central view (CITE: Candidate Imprinted Transcripts by Expression). The database, EICO (Expression-based Imprint Candidate Organizer), can be accessed via the World Wide Web (http://fantom2.gsc.riken.jp/EICODB/) and the DAS client software. These data and interfaces facilitate understanding of the mechanism of imprinting in mammalian inherited traits.

INTRODUCTION

Genomic imprinting results in the expression of individual genes from only one of two parental chromosomes and affects growth and behavior after birth in mammals (1). Aberrant imprinting can lead to various diseases due to an effective doubling of gene dosage. Conversely, genetic diseases display complex inheritance patterns, through the male or female line, when the affected gene falls within a maternally or paternally imprinted locus. Identification of the network of imprinted genes will provide insight into the molecular mechanisms that underlie imprinting-related phenotypes and diseases. To date ~60 imprinted mouse genes have been identified using various methods (http://www.mgu.har.mrc.ac.uk/imprinting/all_impmaps.html). Genomic imprinting involves promoter methylation and/or natural antisense transcripts (NATs) of imprinted or neighboring genes (2); however, the details are unclear. Imprinting clearly cannot be predicted from genomic sequencing and annotation alone (1). We have established an efficient method of screening for candidate imprinted transcripts, and target genes by comparing mRNA expression profiles between parthenogenotes and androgenotes using RIKEN cDNA microarrays (3,4). Although our screening method is very efficient, a fraction (32%) of the identified candidate genes proved to be non-imprinted (3). These non-imprinted genes could be regulated by imprinted genes. To confirm the imprinted status of candidate transcripts, we performed reciprocal crosses with *Mus musculus molossinus* (MSM), a Japanese wild mouse strain, and analyzed the resulting transcripts for polymorphisms that distinguish paternal from maternal loci. Since MSM is phylogenetically 1 million years apart from common laboratory mouse strains, it exhibits frequent genetic polymorphisms with laboratory mice. To this end, we searched for polymorphisms in the 3′-end of the transcripts between MSM and C57BL/6J mouse lines and the results were assembled into the EICO. In this paper, we report the construction and implementation of the EICO (http://fantom2.gsc.riken.jp/EICODB/), which efficiently stores and retrieves three kinds of data: (i) candidate imprinted transcripts from microarray analysis, (ii) single nucleotide polymorphisms (SNPs) between the 3′-end sequences of the RIKEN full-length cDNAs from C57BL/6J and MSM mice, and (iii) imprinting-related disease loci extracted from OMIM (5). The relationship between disease loci and novel imprinted mRNAs identifies new candidates that may be involved causally in imprinting-related human genetic diseases.

DATABASE STRUCTURE AND CONTENTS

The EICO contains 2850 SNPs between C57BL/6J and MSM found in 1281 RIKEN mouse full-length cDNA clones and 2101 candidate imprinted genes derived from microarray experiment data (Table 1). Of the 2101 candidate imprinted genes, 1403 showed maternal expression and 698 showed paternal expression. There were 243 candidate imprinted

*To whom correspondence should be addressed at Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, 1397-1 Yamane, Hidaka City, Saitama 350-1241, Japan. Tel: +81 429 85 7319; Fax: +81 429 85 7329; Email: okazaki@saitama-med.ac.jp

Table 1. Contents of the EICO

Contents	Number
SNPs between MSM and C57BL/6J (1)	2850 (1281 genes)
Candidate imprinted genes (2)	Paternal: 698; maternal: 1403
Genes overlapped between 1 and 2 (3)	243
Candidate imprinted genes on predicted imprinting-related disease loci on human genome (4)	529 genes; 65 diseases; 109 loci
Genes overlapped between 3 and 4	114 genes

The EICO has 2850 SNPs, 2101 candidate imprinted genes and 529 candidate imprinted genes within predicted imprinting-related disease loci.

Figure 1. Web-based interfaces for the EICO. The EICO has two user-friendly web-based interfaces. The CITE system interface is for candidate imprinted genes and imprinting-related diseases. The first color box shows whether the gene is maternally (red) or paternally imprinted (blue). The second box shows whether the gene is mapped to the known imprinted cluster loci. The third, fourth and fifth boxes show whether the gene overlaps with natural antisense transcripts (green), non-coding RNAs (blue) or SNPs (purple). The MuSCAT system can be browsed by SNPs between C57BL/6J and MSM in RIKEN full-length cDNAs, a pair of primer sequences, sequencer name, sequence quality (phred score) and functional annotation of cDNAs. A user can toggle between SNP and candidate gene information through hyperlinks using a typical web browser.

genes included in the 1281 RIKEN mouse full-length cDNA clones. The EICO contains 65 predicted imprinting-related disease (109 disease loci) on the mouse draft genome. A total of 529 candidate imprinted genes extracted from the microarray study were mapped within the disease loci.

The EICO consists of two search interfaces: MoUse SNP CATalog (MuSCAT) and Candidate Imprinted Transcripts by gene Expression (CITE) (Fig. 1). The MuSCAT system retrieves the following information: SNPs, genotype, SNP position on the RIKEN full-length cDNA, sequence quality score (phred score) (6,7), sequence primer pairs, sequencer name, physical position of the cDNAs in the mouse draft genome and functional annotation of the cDNA sequences. MuSCAT links to CITE, FANTOM (http://fantom2.gsc.riken.jp/) (8–10), and EnsEMBL (http://www.ensembl.org/) (11) through hyperlinks and/or DAS (http://www.biodas.org/) (12) on the world wide web. MuSCAT has two user-friendly interfaces: a clone-central view and a SNP-central view. The clone-central view shows the list of all SNPs that were experimentally confirmed on the cDNA clones. The SNP-central view shows the details of the SNP information, such as genotype, sequence quality, sequence primers and the sequencer name.

All candidate imprinted genes can be browsed in the CITE system. The candidate imprinted genes were extracted from the microarray data by comparing mRNA expression in parthenogenotes and androgenotes using RIKEN cDNA microarrays (3,4). The CITE system browses the candidate imprinted genes, their physical position in the mouse and human genomes, the position of imprinting-related disease loci and the functional annotation of these genes. CITE links to OMIM, FANTOM and EnsEMBL using hyperlinks and/or DAS (Fig. 2). CITE has two web interfaces: the candidate imprinted transcripts map view and the disease view. The map view shows the candidate imprinted genes, functional annotation, imprinting status (maternal or paternal) and genomic position of those genes for each chromosome. The disease view can be browsed only from the transcript map view of the human chromosome. The disease view shows the candidate imprinted disease name and information (annotation, map information and imprint status). The candidate gene map information was obtained by mapping mouse candidate imprinted genes on the human genome using *in silico* mapping.

The contents of the EICO will accelerate the discovery of novel disease-related imprinted genes because imprinting is efficiently and quickly confirmed using the RNAs from the reciprocal mouse crosses. To find candidate genes of interest, the EICO can be searched from several biological viewpoints: (i) the presence of NATs (13), (ii) whether the genomic

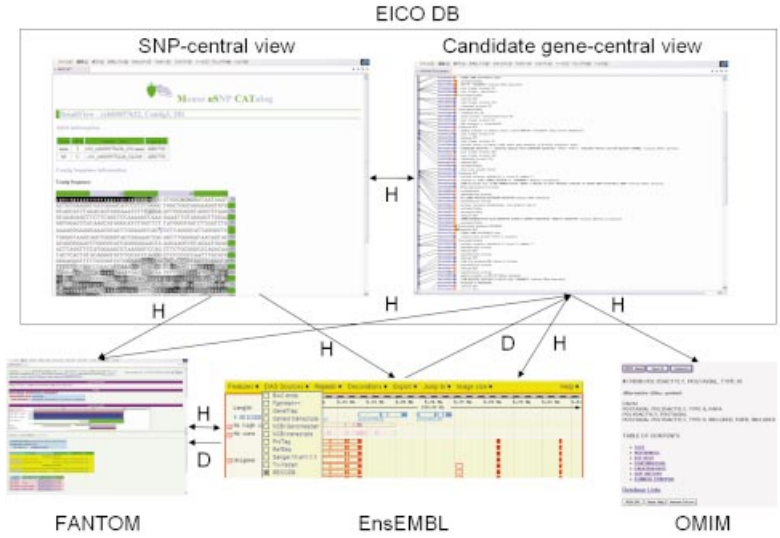

Figure 2. The EICO and public databases. The EICO consists of two searching interfaces: the MuSCAT system for SNP data and the CITE system for candidate imprinted gene data. The EICO links to public databases using hyperlinks and/or DAS. It uses a typical World Wide web browser or DAS client software. H, hyperlink; D, link using DAS.

position of a candidate gene is within an imprinted-related disease locus, (iii) whether the genomic position of a candidate gene is close to a known imprinting cluster and (iv) whether the candidate gene is non-coding RNA (ncRNA) (14). This information can be accessed by the color bar code on the web interface (Fig. 1). The EICO includes 159 NATs, 56 ncRNA and 39 genes mapped to known imprinted cluster loci. Finally, the EICO can be queried with elements such as RIKEN clone ID, RIKEN Rearray ID, FANTOM Annotation, nucleic acid and amino acid sequence using SSAHA (15) and BLAST (16). These data and interfaces in the EICO will serve as a major resource for understanding the mechanism of imprinting in mammalian inherited traits.

IMPLEMENTATION

The EICO is currently implemented using MySQL, an open source relational database management system (http://www.mysql.com/), on Kondara MNU/Linux. MuSCAT. The CITE interface systems are based on an Apache web server (http://www.apache.org/) and CGI programs written in Perl (http://www.cpan.org/) and the object-oriented scripting language Ruby (http://www.ruby-lang.org/). To make hyperlinks to other databases interactive, the EICO uses the DAS protocol using Lightweight Distributed Annotation System (LDAS) (http://www.biodas.org/servers/).

DATA AVAILABILITY AND CITING THE EICO

All users can interactively access all candidate imprinted genes, SNPs and candidate imprinted genes mapped to predicted imprinting-related disease loci via the world wide web at the following URL: http://fantom2.gsc.riken.jp/EICODB/. The MuSCAT and CITE searching systems can be accessed at http://fantom2.gsc.riken.jp/EICODB/snp/, http://fantom2.gsc.riken.jp/EICODB/imprinting/. The sequence similarity search interfaces for the EICO are http://fantom2.gsc.riken.jp/EICODB/ssaha/ and http://fantom2.gsc.riken.jp/EICODB/blast/. The server for the DAS for the EICO services is at http://fantom2.gsc.riken.jp/EICODB/cgi-bin/das/. Please refer to this article and Nikaido *et al.* (4) when citing the EICO.

FUTURE DIRECTIONS

Novel imprinted candidate genes in the EICO will be increased by progressive accumulation of RIKEN full-length cDNA microarray data. The information of validated candidate imprinted genes will be reflected in the EICO when the data are updated. The EICO will import public mouse SNPs within confirmed candidate imprinted genes.

ACKNOWLEDGEMENTS

We thank the following: Yosuke Mizuno, Hidemasa Bono, Shiro Fukuda, Takeya Kasukawa, Ken Yagi, Naoko Tominaga, Yuki Tsujimura, Tomohiro Kono, Yukiko Yamazaki, Toshihiko Shiroishi and Kazuo Moriwaki for technical assistance and discussion. We would like to acknowledge David Hume of the University of Queensland and Elva Diaz of the University of California at Davis for helpful discussion and English editing. This study was also supported by the Special Coordination Fund for the Promotion of Science and Technology; a fund entrusted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to Y.O. and by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H.

REFERENCES

1. Reik,W. and Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.*, **2**, 21–32.
2. Sleutels,F., Zwart,R. and Barlow,D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810–813.

3. Mizuno,Y., Sotomaru,Y., Katsuzawa,Y., Kono,T., Meguro,M., Oshimura,M., Kawai,J., Tomaru,Y., Kiyosawa,H., Nikaido,I. *et al.* (2002) Asb4, Ata3 and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. *Biochem. Biophys. Res. Commun.*, **290**, 1499–1505.
4. Nikaido,I., Saito,C., Mizuno,Y., Meguro,M., Bono,H., Kadomura,M., Kono,T., Morris,G.A., Lyons,P.A., Oshimura,M. *et al.* (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res.*, **13**, 1402–1409.
5. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
6. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
7. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
8. Bono,H., Kasukawa,T., Furuno,M., Hayashizaki,Y. and Okazaki,Y. (2002) FANTOM DB: database of functional annotation of RIKEN mouse cDNA clones. *Nucleic Acids Res.*, **30**, 116–118.
9. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
10. Kasukawa,T., Furuno,M., Nikaido,I., Bono,H., Hume,D.A., Bult,C., Hill,D.P., Baldarelli,R., Gough,J., Kanapin,A. *et al.* (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.*, **13**, 1542–1551.
11. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
12. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
13. Kiyosawa,H., Yamanaka,I., Osato,N., Kondo,S. and Hayashizaki,Y. (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.*, **13**, 1324–1334.
14. Numata,K., Kanai,A., Saito,R., Kondo,S., Adachi,J., Wilming,L.G., Hume,D.A., Hayashizaki,Y. and Tomita,M. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.*, **13**, 1301–1306.
15. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.